Notes on Bernoulli and Binomial random variables

October 1, 2010

1 Expectation and Variance

1.1 Definitions

I suppose it is a good time to talk about expectation and variance, since they will be needed in our discussion on Bernoulli and Binomial random variables, as well as for later disucssion (in a forthcoming lecture) of Poisson processes and Poisson random variables.

Basically, given a random variable $X : \mathcal{S} \to \mathbb{R}$, having a pdf $f(x)$, define the expectation to be

$$\mathbb{E}(X) \ := \ \int_{-\infty}^{\infty} x f(x) dx.$$

In other words, it is kind of like the "average value of X weighted by $f(x)$". If X is discrete, where say X can take on any of the values $a_1, a_2, ...$, we would have

$$\mathbb{E}(X) \ := \ \sum_i a_i p(a_i),$$

where p here denotes the mass function.

More generally, if $\nu : \mathcal{B}(\mathbb{R}) \to [0, 1]$ is a probability measure associated to a random variable X (continuous, discrete, or otherwise), so that for $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbb{P}(X \in A) \ = \ \int_{\mathbb{R}} d\nu(x) \ = \ \nu(A),$$

we define

$$\mathbb{E}(X) := \int_{\mathbb{R}} x d\nu(x).$$

Of course, we really haven't worked much with the sort of r.v.'s for which such a ν cannot be relized in terms of pdf's, but I thought I would point it out anyways.

We likewise define the variance operator $V(X)$ to be

$$V(X) := \int_{\mathbb{R}} (x-\mu)^2 f(x) dx, \text{ where } \mu = \mathbb{E}(X).$$

If X is discrete, then

$$V(X) := \sum_i (a_i - \mu)^2 p(a_i);$$

and, of course, there is the even more general version

$$V(X) := \int_R (x-\mu)^2 d\mu(x).$$

We know that the expectation is a kind of average, and now I want to give you a feel for what the variance is: Basically, it is a measure of how "flat" the pdf is – the flatter it is, the more the values of X away from the mean μ get weighted; and therefore the larger the variance $V(X)$ is. For example, consider the two random variables X and Y having pdf's $f(x)$ and $g(x)$, respectively, given by

$$f(x) = \begin{cases} 1/2, & \text{if } x \in [-1,1]; \\ 0, & \text{if } x < -1 \text{ or } x > 1. \end{cases}$$

and

$$g(x) = \begin{cases} 1/4, & \text{if } x \in [-2,2]; \\ 0, & \text{if } x < -2 \text{ or } x > 2. \end{cases}$$

In some sense, $g(x)$ is "flatter" than $f(x)$, since the total mass of 1 (the integral of $f(x)$ and $g(x)$ over the whole real line) gets spread out over an interval of width 4 in the case of $g(x)$, but only over an interval of width 2 in the case of $f(x)$. So, the variance of Y should be larger than the variance of X. Let's check our intuition: First, it is a simple calculation to show that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, so that

$$V(X) \ = \ \int_{-1}^{1} (x-0)^2 f(x) dx \ = \ \int_{-1}^{1} (x^2/2) dx \ = \ x^3/6 \Big|_{-1}^{1} \ = \ 1/3;$$

and,

$$V(Y) \ = \ \int_{-2}^{2} (y-0)^2 g(y) dy \ = \ \int_{-2}^{2} (y^2/4) dy \ = \ y^3/12 \Big|_{-2}^{2} \ = \ 4/3.$$

Our guess was indeed correct.

You may wonder to yourself: Why don't we define the variance as just $\mathbb{E}(|X-\mu|)$, instead of $\mathbb{E}((X-\mu)^2)$? There are several reasons for this. One reason is that $(x-\mu)^2$ is differentiable, and therefore can be analyzed using the methods of calculus; whereas, $|x-\mu|$ is not. And another reason is that there are certain nice identities for computing the variance when squares are used that lend themselves to something called the "second moment method", which we will discuss in due course.

1.2 The mean and variance may not even exist!

There are some random variables that are so spread out that even the expected value $\mathbb{E}(X)$ does not exist. For example, suppose that X has the pdf

$$f(x) \ = \ \begin{cases} 1/(2x^{3/2}), & \text{if } x \geq 1; \\ 0, & \text{otherwise.} \end{cases}$$

First, let's confirm it is a valid pdf:

$$\int_{\mathbb{R}} f(x) dx \ = \ \int_{1}^{\infty} dx/(2x^{3/2}) \ = \ x^{-1/2} \Big|_{1}^{\infty} \ = \ 1.$$

And now, for the expectation we get

$$\mathbb{E}(X) \ = \ \int_{1}^{\infty} 1/(2\sqrt{x}) dx \ = \ \infty.$$

2 Bernoulli and Binomial random variables

A Bernoulli random variable X is one that takes on the values 0 or 1 according to

$$\mathbb{P}(X = j) \ = \ \begin{cases} p, & \text{if } j = 1; \\ q = 1 - p, & \text{if } j = 0. \end{cases}$$

Things only get interesting when one adds several independent Bernoulli's together. First, though, we need a notion of independent random variables: We say that a collection of random variables

$$X_1, ..., X_n \ : \ \mathcal{S} \ \to \ \mathbb{R}$$

are *independent* if for every sequence $A_1, ..., A_n \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbb{P}(X_1 \in A_1, \ X_2 \in A_2, \ ..., \ X_n \in A_n) \ = \ \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n).$$

This may seem a little strange, since before when we defined independent events we considered all possible subsequences $1 \leq i_1 < \cdots < i_k \leq n$ of indices when intersecting our events, whereas here we work with *all* the random variables. But, in fact, we have gotten around this issue in the above, because by letting some of the $A_i = \mathbb{R}$ itself, we get that $X_i \in A_i = \mathbb{R}$ conveys no information; so, if in fact we have that $A_i = \mathbb{R}$ for $i \neq i_1, ..., i_k$, then

$$\mathbb{P}(X_1 \in A_1, \ ..., \ X_n \in A_n) \ = \ \mathbb{P}(X_{i_1} \in A_{i_1}, \ ..., \ X_{i_k} \in A_{i_k}).$$

Now we arrive at the following claim:

Claim. Let $X = X_1 + \cdots + X_n$, where the X_i's are independent Bernoulli random variables having the same parameter p. Then, X has a Binomial distribution with parameters n and p; that is,

$$\mathbb{P}(X = x) \ = \ \binom{n}{x} p^x q^{n-x}, \ \text{where } q = 1 - p.$$

To see this, let us begin by letting $\mathcal{S} = \{(t_1, ..., t_n) \ : \ t_i = 0 \text{ or } 1\}$; that is, $\mathcal{S}$ is the sample space associated with the collection of random variables, where t_i corresponds to the value of X_i; so, the event $X_1 = 1$ corresponds to the 2^{n-1} vectors

$$(1, 0, 0, ..., 0),\ (1, 1, 0, ..., 0),\ (1, 0, 1, 0, ..., 0),\ (1, 1, 1, 0, ..., 0),\ ...$$

Basically, all those vectors with first coordinate 1.

We just let $\Sigma = 2^{\mathcal{S}}$, and then the probability measure we put on Σ is built up from the measure applied to singletons via

$$\mathbb{P}(\{(t_1, ..., t_n)\} \ = \ p^{t_1+\cdots+t_n} q^{n-t_1-\cdots-t_n}.$$

One can check that this is exactly the measure assigning the events $X_i = 1$ probability p, while keeping the X_i's independent random variables.

Now, the event $E \in \Sigma$ corresponding to $X_1 + \cdots + X_n = x$ is made up out of $\binom{n}{x}$ singletons, where all these singletons have the same probability $p^x q^{n-x}$; and so, we are led to

$$\mathbb{P}(E) \ = \ \binom{n}{x} p^x q^{n-x},$$

by writing E as a disjoint union of singletons.

3 Facts about Binomial random variables

If X is a Binomial random variable with parameters n and p, then

$$\mathbb{E}(X) \ = \ np,\ V(X) \ = \ npq.$$

A very simple way we could show this is to use something called the "linearity of expectation", along with the fact that $X \sim X_1 + \cdots + X_n$, where the notation $X \sim Y$ means that X and Y have the same distribution. However, this requires some facts about multi-dimensional random variables that we have not yet covered, and so we will take a more low-brow approach for now. Basically, I will give two different low-tech ways to prove the identity above for $\mathbb{E}(X)$; the identity for $V(X)$ follows from similar, though more involved, manipulations.

3.1 Way 1: The derivative trick

Treat p as some varialbe, and let q be a constant which may or may not depend on p (i.e. we suspend using the identity $q = 1 - p$). Then, we are led to the function

$$h(p) \ = \ \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}.$$

Now take a derivative with respect to p, so as to introduce an extra factor of j into each term in the sum; that is,

$$h'(p) \ = \ \frac{dh(p)}{dp} \ = \ \sum_{j=0}^{n} \binom{n}{j} j p^{j-1} q^{n-j}.$$

Inserting an extra factor of p into each term, to give

$$ph'(p) \ = \ \sum_{j=0}^{n} \binom{n}{j} j p^j q^{n-j},$$

we see that this is the expectation of a Binomial r.v. with parameters n and p.

Now, by the binomial theorem we know that

$$(p+q)^n \ = \ h(p),$$

from which we deduce

$$h'(p) \ = \ n(p+q)^{n-1}.$$

At this point we now revert to thinking of $q = 1 - p$, so that

$$h'(p) \ = \ n;$$

and therefore that the expectation of our Binomial r.v. is np, as claimed.

3.2 Way 2: Rewrite the binomial coefficient, and reinterpret

Observe that

$$j\binom{n}{j} \ = \ n\binom{n-1}{j-1}.$$

So, if X is our Bernoulli r.v., we have that

$$\mathbb{E}(X) \;=\; \sum_{j=0}^{n} \binom{n}{j} j p^j q^{n-j} \;=\; \sum_{j=1}^{n} \binom{n}{j} j p^j q^{n-j} \;=\; np \sum_{j=1}^{n} \binom{n-1}{j-1} p^{j-1} q^{(n-1)-(j-1)}.$$

Renumbering by seeting $k = j - 1$ gives

$$\mathbb{E}(X) \;=\; np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{(n-1)-k}.$$

And now, one may either interpret this last sum probabilistically in terms of a Binomial r.v. with parameters $n-1$ and p, or one can just apply the Binomial theorem – either way, we see that it equals 1, giving $\mathbb{E}(X) = np$ once again.